

Malignant Comments Classifier

Submitted by:-
Mohit Meel

Acknowledgement

- The data is secondary as it is provided by the company, it is not collected from anywhere.
- The company has provided all the guidelines to complete the project as the company has provided to problem statement, data description, pointers files related to project information .
- The data is in csv format and has two files I.e train file and test file.
- The data is secondary therefore no researches has been made.
- The company has provided sample documentation and description of data in order to complete the project and also to study the project.



Introduction

- The problem statement is to predict the comments given by the public is malignant or not. In today's world, the social websites or some other platforms such as Facebook, Instagram, you tube, twitter, etc., where people share their feelings by uploading photos, text, etc. In this social websites, the other people write their text as comments on the people photos, text, etc. The comment written is can be malignant or not. In other words, there are some people who writes negative comments on the people photos, etc.
- The research done on this project done is that I personally visited to the social websites, I came to know that there are so many people who writes negative comments on the people sharing their photos which results to those people in having depression or feeling hurt.
- The objective of working on this project is that these peoples should be caught by the cyber police because most of the cases related suicide because of depression due to malignant comments is coming now a days. These comments written should be stopped.

Analytical Problem Framing

- The data sources is the secondary as it is provided by the company. The data is in csv and has two files: train and test file . It is huge data as train file has 159000 samples and test file has 153000 samples. The train file is for finding out the best model and the test file to predict the target variable. The data is object type. There are 8 columns in the train file and two columns in the test file.
- The data is first analyzed and understand by writing the code such as head, tail, dtypes, columns, info, describe, Corr, shape, skew, is null, then the data cleaning process is done in which the comment text column of both the files has first converted into lower case next the replaced/removed the email address, URL with web address, money symbols, phone numbers and numbers if any after that the ignoring of punctuation, stop words is done next using word net lemmatize next using the tf-idf vectorizer for converting the text into vectors.
- The relationship between the input and output is that the input (that is comment text which in the text format) and the output (that is in binary digit I.e. 0 and 1 which shows 0 is no and 1 is yes) is highly correlated as these comments will be taken to ascertain whether these are malignant or not. The output variables are malignant, highly malignant, rude, threat, abuse, loathe. I have done sum of all the columns of output variables and then made a separate column of name target which denotes the target variable.
- The only assumption I have taken that the target column of is showing class imbalance so I have removing class imbalance using smote technique and the datatype of data is object therefore I have not removed outliers and skewness present in it.
- The libraries that are used is pandas as seaborn, matplotlib, warnings, NumPy, word tokenize, reg exp tokenize, stop words, wordnet, string, Word Net Lemmatize, Porter Stemmer.

Model Development and Evaluation

- The approaches I have made for solving the problem is classification approach as the prediction variable is discrete which mean only two output will be there that is 0 or 1. I have send the 70% of train file for training and 30% of train file for testing and random state is 45.
- The models which I have taken for the training and testing is logistic regression, decision tree classifiers, random forest classifiers, ada boost classifiers, gradient boosting classifiers.
- The models listed above has the accuracy score of 92.9%, 93.8%, 97.4%, 80.6% and 83% respectively. The best which comes out from all listed above models is the random forest classifiers with an accuracy score of 97.4%.
- The key metrics used are accuracy score, confusion matrix, classification report, roc auc score, roc curve and grid search cv for the best model.
- The visualization part done on the project is the heatmap for knowing null values and correlation between the input and output variable, the boxplot for knowing the outliers present, the kde-plot, hist-plot for checking the skewness in the dataset and the roc curve plot showing the roc auc score of random forest classifiers.
- The results come up from visualization is that the outliers are present and skewness is also present but can't be removed because of object datatype if removed then data will become biased. The roc curve is exactly it should show according to the accuracy score achieved from random forest classifier

Conclusion

- The key findings and observation from the problem is that the comments are malignant because of the nature of the people. How they see the people on their view which directly shows their thinking is bad in nature which is quite dangerous for the society. Most of the malignant comments are on the girls photos through the male people which I have observed on the social websites which shows that how such male people has down to earth thinking.
- The learning for the process of visualization is discussed in the previous slide but the learning from the data cleaning is that it had a very good experience to knowing and learning new NLP techniques of cleaning text related data such stop words, punctuations, vectorization, etc. The best model that comes is random forest classifiers which I have used to predict the target variable of test file as the accuracy score is 97.4%. I faced many challenges as the NLP techniques were new for me because I have never learned about NLP with machine learning and not taught too.
- According to me I think there is only one limitation which is that I have analyzed the predicted values of test file which I came to know that some of the comments are not malignant but it shows 1 which means yes it is malignant. To improve the results the comment text of train file have accuracy in it.